HarmonyView: Harmonizing Consistency and Diversity in One-Image-to-3D

Sangmin Woo[1*] Byeongjun Park[1*] Hyojun Go[2] Jin-Young Kim[2] Changick Kim[1]
[1]KAIST [2]Twelve Labs
[1]{smwoo95, pbj3810, changick}@kaist.ac.kr [2]{william, jeremy}@twelvelabs.io

https://byeongjun-park.github.io/HarmonyView/

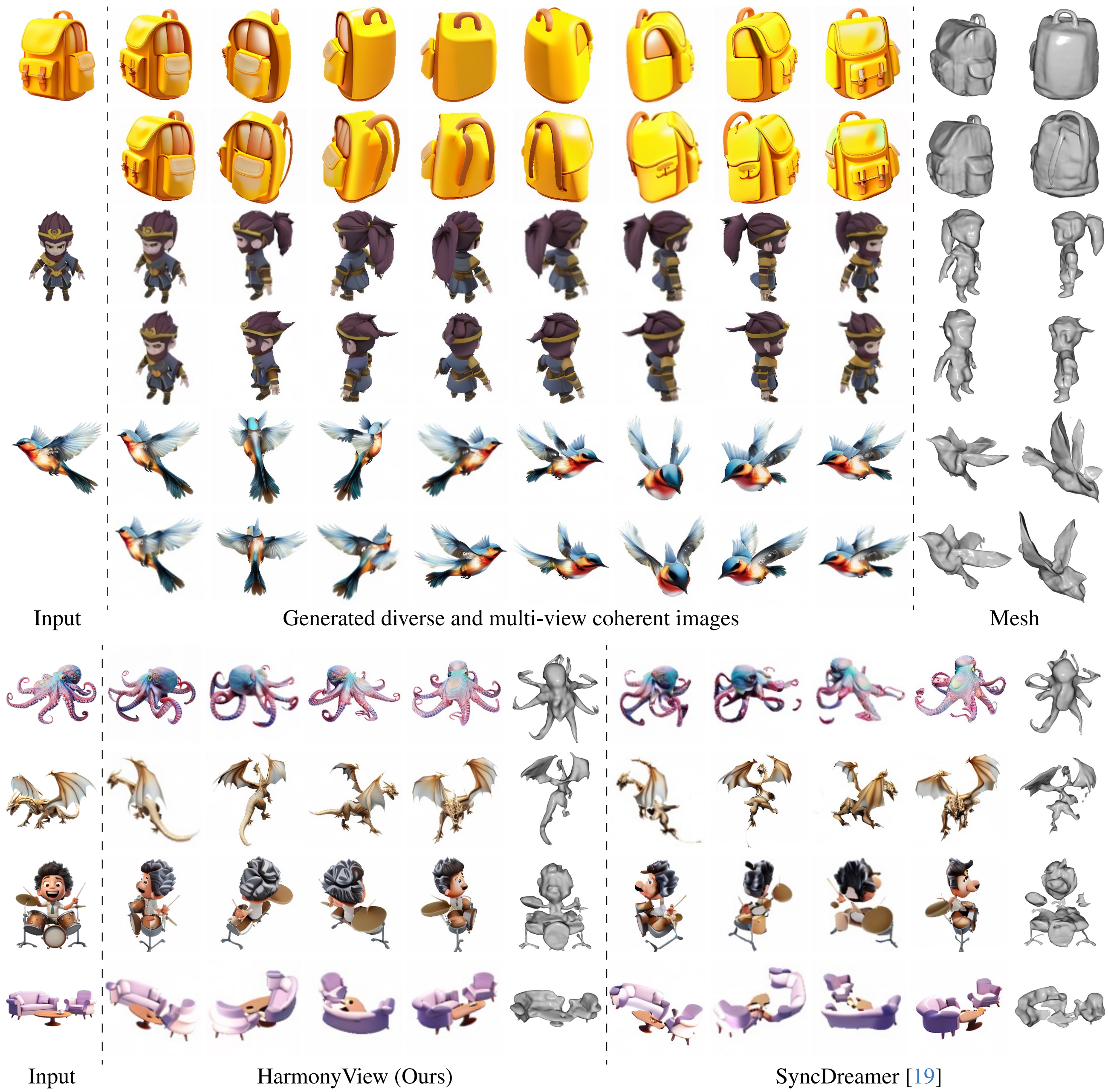

Figure 1. **HarmonyView for one-image-to-3D.** HarmonyView generates realistic 3D content using just a single image. It excels at maintaining visual and geometric consistency across generated views while enhancing the diversity of novel views, even in complex scenes.

*Equal contribution

Abstract

Recent progress in single-image 3D generation highlights the importance of multi-view coherency, leveraging 3D priors from large-scale diffusion models pretrained on Internet-scale images. However, the aspect of novel-view diversity remains underexplored within the research landscape due to the ambiguity in converting a 2D image into 3D content, where numerous potential shapes can emerge. Here, we aim to address this research gap by simultaneously addressing both consistency and diversity. Yet, striking a balance between these two aspects poses a considerable challenge due to their inherent trade-offs. This work introduces ***HarmonyView****, a simple yet effective diffusion sampling technique adept at decomposing two intricate aspects in single-image 3D generation: consistency and diversity. This approach paves the way for a more nuanced exploration of the two critical dimensions within the sampling process. Moreover, we propose a new evaluation metric based on CLIP image and text encoders to comprehensively assess the diversity of the generated views, which closely aligns with human evaluators' judgments. In experiments, HarmonyView achieves a harmonious balance, demonstrating a win-win scenario in both consistency and diversity.*

1. Introduction

Humans can effortlessly imagine the 3D form of an object from just a single camera view, drawing upon their prior knowledge of the 3D world. Yet, emulating this human capability in machines remains a longstanding challenge in the field of computer vision [2, 26, 37, 44, 46, 59]. The fundamental hurdle lies in the inherent ambiguity of deducing 3D structure from a single 2D image since a single image essentially collapses the three dimensions of the real world into a 2D representation. Consequently, countless 3D configurations of an object can be projected onto the same 2D image. This ambiguity has ignited the quest for innovative solutions for single-image 3D generation [1, 12, 14, 16, 17, 19, 20, 29, 33–36, 40–42, 51, 52, 56, 57, 60, 61].

One prevalent strategy is to generate multi-view images from a single 2D image [17, 18, 40, 50], and process them using techniques such as Neural Radiance Fields (NeRFs) [23] to create 3D representations. Regarding this, recent studies [18, 19, 40, 50, 56, 57] highlight the importance of maintaining *multi-view coherency*. This ensures that the generated 3D objects to be coherent across diverse viewpoints, empowering NeRF to produce accurate and realistic 3D reconstructions. To achieve this, researchers harness the capabilities of large-scale diffusion models [32], particularly those trained on a vast collection of 2D images. The abundance of 2D images provides a rich variety of views for the same object, allowing the model to learn view-to-view relationships and acquire geometric priors about the 3D world. On top of this, some works [19, 40] introduce a refinement stage that fine-tunes the view alignment to accommodate variations in camera angles. This adjustment is a key factor in achieving the desired multi-view coherency, which directly impacts the realism of the resulting 3D representation. This progress has notably enhanced the utility of the generated 3D contents, making them more suitable for various applications [28, 53].

An equally significant but often overlooked aspect in single-image 3D generation is the *novel-view diversity*. The ill-posed nature of this task necessitates dealing with numerous potential 3D interpretations of a given 2D image. Recent works [18, 19, 40, 49] showcase the potential of creating diverse 3D contents by leveraging the capability of diffusion models in generating diverse 2D samples. However, balancing the pursuit of consistency and diversity remains a challenge due to their inherent trade-off: maintaining *visual consistency* between generated multi-view images and the input view image directly contributes to sample quality but comes at the cost of limiting *diversity*. Although current multi-view diffusion models [19, 40] attempt to optimize both aspects simultaneously, they fall short of fully unraveling their intricacies. This poses a crucial question: Can we navigate towards a harmonious balance between these two fundamental aspects in single-image 3D generation, thereby unlocking their full potential?

This work aims to address this question by introducing a simple yet effective diffusion sampling technique, termed HarmonyView. This technique effectively decomposes the intricacies in balancing consistency and diversity, enabling a more nuanced exploration of these two fundamental facets in single-image 3D generation. Notably, HarmonyView provides a means to exert explicit control over the sampling process, facilitating a more refined and controlled generation of 3D contents. This versatility of HarmonyView is illustrated in Fig. 1. Our method achieves a harmonious balance, demonstrating mutual benefits in both consistency and diversity. HarmonyView generates geometrically coherent 3D contents that faithfully represent the input image for visible parts while also capturing diverse yet plausible modes for occluded parts. Another challenge we face is the absence of standardized metrics for assessing the diversity of generated multi-views. To address this gap and provide a more comprehensive assessment of the consistency and diversity of 3D contents, we introduce a novel evaluation metric based on both the CLIP image and text encoders [8, 30].

In experiments, we quantitatively compare HarmonyView against state-of-the-art techniques, spanning two tasks: novel-view synthesis and 3D reconstruction. In both tasks, HarmonyView consistently outperforms baseline methods across all metrics. Our qualitative results further highlight the efficacy of HarmonyView, showcasing faithful reconstructions with remarkable visual quality, even in complex scenes.

Moreover, we show that our proposed metric closely aligns with the assessments made by human evaluators. Lastly, HarmonyView can be seamlessly integrated with off-the-shelf text-to-image diffusion models (*e.g*., Stable Diffusion [32]), enabling it to perform text-to-image-to-3D generation.

2. Related Work

Lifting 2D pretrained models for 3D generation. Recent research endeavors [3, 15, 21, 36, 42, 45, 49, 52, 61] are centered on the idea of lifting 2D pre-trained models [30, 32] to create 3D models from textual prompts, without the need for explicit 3D data. The key insight lies in leveraging 3D priors acquired by diffusion models during pre-training on Internet-scale data. This enables them to *dream up* novel 3D shapes guided by text descriptions. DreamFusion [27] distills pre-trained Stable Diffusion [32] using *Score Distillation Sampling (SDS)* to extract a Neural Radiance Field (NeRF) [23] from a given text prompt. DreamFields [11] generates 3D models based on text prompts by optimizing the CLIP [30] distance between the CLIP text embedding and NeRF [23] renderings. However, accurately representing 3D details with word embeddings remains a challenge.

Similarly, some works [22, 29, 41, 55] extend the distillation process to train NeRF for the 2D-to-3D task. NeuralLift-360 [55] utilizes a depth-aware NeRF to generate scenes guided by diffusion models and incorporates a distillation loss for CLIP-guided diffusion prior [30]. Magic123 [29] uses SDS loss to train a NeRF and then fine-tunes a mesh representation. Due to the reliance on SDS loss, these methods necessitate textual inversion [7] to find a suitable text description for the input image. Such a process needs per-scene optimization, making it time-consuming and requiring tedious parameter tuning for satisfactory quality.

Another line of work [17, 18, 40, 50] uses 2D diffusion models to generate multi-view images then use them for 3D reconstruction with NeRF [23, 47]. 3DiM [50] views novel-view synthesis as an image-to-image translation problem and uses a pose-conditional diffusion model to predict novel views from an input view. Zero-1-to-3 [18] enables zero-shot 3D creation from arbitrary images by fine-tuning Stable Diffusion [32] with relative camera pose. Our work, falling into this category, is able to convert arbitrary 2D images to 3D without SDS loss [27]. It seamlessly integrates with other frameworks, such as text-to-2D [24, 31, 32] and neural reconstruction methods [23, 47], streamlining the text-to-image-to-3D process. Unlike prior distillation-based methods [22, 55] confined to a singular mode, our approach offers greater flexibility for generating diverse 3D contents.

Consistency and diversity in 3D generation. The primary challenge in single-image 3D content creation lies in maintaining multi-view coherency. Various approaches [18, 19, 50, 56, 57] attempt to tackle this challenge: Viewset Diffusion [40] utilizes a diffusion model trained on multi-view 2D data to output 2D viewsets and corresponding 3D models. SyncDreamer [19] introduces a 3D-aware feature attention that synchronizes intermediate states of noisy multi-views. Despite these efforts, achieving complete geometric coherence in generated views remains a challenge.

On the other hand, diversity across generated 3D samples is another critical aspect in single-image 3D generation. However, only a few works in the related literature specifically address this issue, often limited to domains such as face generation [4] or starting from text for 3D generation [49]. Recent studies [18, 19, 40, 57] showcase the potential of pre-trained diffusion models [32] in generating diverse multi-view images. However, there is still significant room for exploration in balancing consistency and diversity. In our work, we aim to unlock the potential of diffusion models, allowing for reasoning about diverse modes for novel views while being faithful to the input view for observable parts. We achieve this by breaking down the formulation of multi-view diffusion model into two fundamental aspects: *visual consistency* with input view and *diversity* of novel views. Additionally, we propose the CD score to address the absence of a standardized diversity measure in existing literature.

3. Method

Our goal is to create a high-quality 3D object from a single input image, denoted as $\mathbf{y}$. To achieve this, we use the diffusion model [39] to generate a cohesive set of N views at pre-defined viewpoints, denoted as $\mathbf{x}_0^{(1:N)} = \{\mathbf{x}_0^{(1)}, ..., \mathbf{x}_0^{(N)}\}$. These mutli-view images are then utilized in NeRF-like techniques [23, 47] for 3D reconstruction. The key to a realistic 3D object lies in the consistency across the generated views. If they exhibit coherent appearance and geometry, the resulting 3D object will appear more natural. Therefore, ensuring consistency is crucial for achieving our goal. Recent works [19, 34, 40] address multi-view generation by jointly optimizing the distribution of multiple views. Building upon them, we aim to enhance both consistency and diversity by decomposing their formulation during diffusion sampling.

3.1. Multi-view Diffusion Models

We address the challenge of generating a 3D representation from a single, partially observed image using diffusion models [38, 39]. These models inherently possess the capability to capture diverse modes [54], making them well-suited for the task. In the context of multi-view image generation, SyncDreamer [19] introduces a multi-view diffusion model that captures the *joint distribution* of N novel views $\mathbf{x}_0^{(1:N)}$ given an input view $\mathbf{y}$. This model extends the DDPM [10] forward process by adding random noises independently to each view at every time step:

$$\mathbf{x}_t^{(n)} = \sqrt{\bar{\alpha}_t}\mathbf{x}_0^{(n)} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}^{(n)}. \tag{1}$$

Here, n denotes the view index. A noise prediction model $\boldsymbol{\epsilon}_\theta$ is then used to "undo" the forward steps to recover the original data. This model predicts the noise of the n-th view $\boldsymbol{\epsilon}^{(n)}$, given the condition of an input view $\mathbf{y}$, the view difference between the input view and the n-th target view $\Delta\mathbf{v}^{(n)}$, and noisy multi views $\mathbf{x}_t^{(1:N)}$. Hereafter, we define the pair $(\mathbf{y}, \Delta\mathbf{v}^{(n)})$ as the reference view condition $\mathbf{r}^{(n)}$, and the model is trained by the noise prediction objective as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0^{(1:N)}, \boldsymbol{\epsilon}^{(1:N)}, t} \|\boldsymbol{\epsilon}^{(n)} - \boldsymbol{\epsilon}_\theta(\mathbf{x}^{(n)}; t, \mathbf{c}^{(n)})\|_2^2, \quad (2)$$

where $\mathbf{c}^{(n)} = (\mathbf{r}^{(n)}, \mathbf{x}_t^{(1:N)})$ and $\boldsymbol{\epsilon}^{(1:N)}$ represents Gaussian noise of size $N \times H \times W$ added to all N views.

3.2. HarmonyView

Diffusion sampling guidance. Classifier-guided diffusion [5] uses a noise-robust classifier $p(\boldsymbol{l}|\mathbf{x}_t)$, which estimates the class label $\boldsymbol{l}$ given a noisy sample $\mathbf{x}_t$, to guide the diffusion process with gradients $\nabla_{\mathbf{x}_t} \log p(\boldsymbol{l}|\mathbf{x}_t)$. This classifier requires bespoke training to cope with high noise levels (where timestep t is large) and to provide meaningful signals all the way through the sampling process. Classifier-free guidance [9] uses a single conditional diffusion model $p_\theta(\mathbf{x}|\boldsymbol{l})$ with conditioning dropout, which intermittently replaces $\boldsymbol{l}$ (typically 10%) with a null token ϕ (representing the absence of conditioning information) for unconditional predictions. This models an *implicit classifier* directly from a diffusion model without the need for an extra classifier trained on noisy input. These conditional diffusion models [5, 9] dramatically improve sample quality by enhancing the conditioning signal but with a trade-off in diversity.

What's wrong with multi-view diffusion sampling? From Eq. (2), we derive an *unconditional* diffusion model $p(\mathbf{x}^{(n)})$ parameterized by a score estimator $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)}; t)$ and *conditional* diffusion model $p(\mathbf{x}^{(n)}|\mathbf{c}^{(n)})$ parameterized by $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)}; t, \mathbf{c}_t^{(n)})$. These two models are learned via a single neural network following the classifier-free guidance [9]. During sampling, the multi-view diffusion model adjusts its prediction as follows (t is omitted for clarity):

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t^{(n)}; \mathbf{c}^{(n)}) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)}; \mathbf{c}^{(n)}) + s \cdot (\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)}; \mathbf{c}^{(n)}) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)})), \quad (3)$$

where s represents a guidance scale. The model output is extrapolated further in the direction of $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)}; \mathbf{c}_t^{(n)})$ and away from $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)})$. Thus, the scaling of s affects both the input view condition $\mathbf{r}^{(n)}$ and the multi-view condition $\mathbf{x}_t^{(1:N)}$ simultaneously. As evidenced by Fig. 2 and Table 5, increasing s encourages *multi-view coherency* and *diversity* in the generated views. Yet, this comes with a trade-off: it diminishes the *visual consistency* with the input view. While the inherent trade-off between these two dimensions is obvious in this context, managing competing objectives under a single guidance poses a considerable challenge. In essence, the model tends to generate diverse and geometrically coherent multi-view images, but differ in visual aspects (*e.g.*, color, texture) from the input view, resulting in sub-optimal quality.

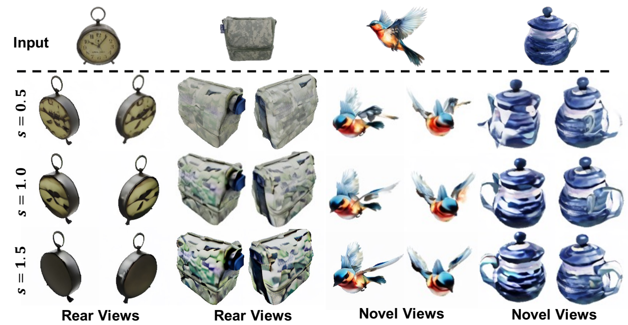

Figure 2. **Visual ablation on scale s of Eq. (3).**

Harmonizing consistency and diversity. To address the aforementioned challenge, we introduce a method termed "HarmonyView". Our approach leverages two implicit classifiers. One classifier $p^i(\mathbf{r}^{(n)}|\mathbf{x}_t^{(n)}, \mathbf{x}_t^{(1:N)})$ guides the target view $\mathbf{x}_t^{(n)}$ and multi-views $\mathbf{x}_t^{(1:N)}$ to be more *visually consistent* with the input view $\mathbf{r}^{(n)}$. Another classifier $p^i(\mathbf{x}_t^{(1:N)}|\mathbf{x}_t^{(n)}, \mathbf{r}^{(n)})$ contains uncertainty in both the target $(\mathbf{x}_t^{(1:N)})$ and conditional $(\mathbf{x}_t^{(n)})$ elements. This contributes to capturing *diverse* modes. Together, they synergistically guide the synchronization of noisy multi-views $\mathbf{x}_t^{(1:N)}$, facilitating *geometric coherency* among clean multi-views. Based on these, we redefine the score estimation as follows:

$$\begin{aligned} \tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t^{(n)}; \mathbf{c}^{(n)}) &= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)}; \mathbf{c}^{(n)}) \\ &- s_1 \sigma_t \nabla_{\mathbf{x}_t^{(n)}} \log p^i(\mathbf{r}^{(n)}|\mathbf{x}_t^{(n)}, \mathbf{x}_t^{(1:N)}) \\ &- s_2 \sigma_t \nabla_{\mathbf{x}_t^{(n)}} \log p^i(\mathbf{x}_t^{(1:N)}|\mathbf{x}_t^{(n)}, \mathbf{r}^{(n)}), \end{aligned} \quad (4)$$

where s_1 and s_2 are guidance scales and σ_t is a noise scheduling parameter. By properly balancing these terms, we can obtain multi-view coherent images that align well with the semantic content of the input image while being diverse across different samples.

According to Bayes' rule, $p^i(\mathbf{r}^{(n)}|\mathbf{x}_t^{(n)}, \mathbf{x}_t^{(1:N)}) \propto p(\mathbf{x}_t^{(n)}|\mathbf{c}^{(n)})/p(\mathbf{x}_t^{(n)}|\mathbf{x}_t^{(1:N)})$ and $p^i(\mathbf{x}_t^{(1:N)}|\mathbf{x}_t^{(n)}, \mathbf{r}^{(n)}) \propto p(\mathbf{x}_t^{(n)}|\mathbf{c}^{(n)})/p(\mathbf{x}_t^{(n)}|\mathbf{r}^{(n)})$. Hence, the diffusion scores of these two implicit classifiers can be derived as follows:

$$\begin{aligned} &\nabla_{\mathbf{x}_t^{(n)}} \log p^i(\mathbf{r}^{(n)}|\mathbf{x}_t^{(n)}, \mathbf{x}_t^{(1:N)}) \\ &\quad = -\frac{1}{\sigma_t}(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)}; \mathbf{c}^{(n)}) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)}; \mathbf{x}_t^{(1:N)})). \end{aligned} \quad (5)$$

$$\begin{aligned} &\nabla_{\mathbf{x}_t^{(n)}} \log p^i(\mathbf{x}_t^{(1:N)}|\mathbf{x}_t^{(n)}, \mathbf{r}^{(n)}) \\ &\quad = -\frac{1}{\sigma_t}(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)}; \mathbf{c}^{(n)}) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(n)}; \mathbf{r}^{(n)})). \end{aligned} \quad (6)$$

Figure 3. **Qualitative comparison of several instantiations for multi-view diffusion guidance on novel-view synthesis.** Our decomposition of Eq. (3) yields two guidance parameters: s_1 for input-target visual consistency and s_2 for diversity in the novel views. With these parameters, our final formulation Eq. (7) enables the generation of a diverse set of multi-view coherent images that well reflect the input view.

Method	s	s_1	s_2	PSNR↑	SSIM↑	LPIPS↓	E_{flow}↓	CD↑
No Guidance				20.51	0.818	0.144	2.270	0.640
Baseline (Eq. (3))	✓			20.19	0.819	0.140	2.071	0.717
			✓	20.32	0.822	0.141	2.136	0.764
Ours (Eq. (7))		✓		**21.03**	**0.828**	**0.128**	2.146	0.668
		✓	✓	20.69	0.825	0.133	**1.945**	**0.792**

Table 1. **Ablative study of multi-view diffusion guidance on novel-view synthesis.** Metrics measure sample quality with PSNR, SSIM, LPIPS; multi-view coherency with E_{flow}; and diversity with CD score. Our final design strikes the best balance across the metrics. Here, we set $s = 1$, $s_1 = 2$, $s_2 = 1$.

Finally, these terms are plugged into Eq. (4) and yields:

$$\begin{aligned}\tilde{\epsilon}_\theta(\mathbf{x}_t^{(n)};\mathbf{c}^{(n)}) &= \epsilon_\theta(\mathbf{x}_t^{(n)};\mathbf{c}^{(n)}) \\ &+ s_1 \cdot (\epsilon_\theta(\mathbf{x}_t^{(n)};\mathbf{c}^{(n)}) - \epsilon_\theta(\mathbf{x}_t^{(n)};\mathbf{x}_t^{(1:N)}) \\ &+ s_2 \cdot (\epsilon_\theta(\mathbf{x}_t^{(n)};\mathbf{c}^{(n)}) - \epsilon_\theta(\mathbf{x}_t^{(n)};\mathbf{r}^{(n)}).\end{aligned} \tag{7}$$

This formulation effectively decomposes *consistency* and *diversity*, offering a nuanced approach that grants control over both dimensions. While simple, our decomposition achieves a win-win scenario, striking a harmonious balance in generating samples that are both consistent and diverse (see Fig. 3 and Table 1).

3.3. Consistency-Diversity (CD) Score

We propose the CD score with two key principles: (1) *Diversity of novel views*: It is preferable that the generated images exhibit diverse and occasionally creative appearances that are not easily imaginable from the input image. (2) *Semantic consistency*: While pursuing diversity, it is crucial to maintain semantic consistency, *i.e.*, the generated images should retain their semantic content consistently, regardless of variations in the camera viewpoint. To operationalize this evaluation, CD score utilizes CLIP [30] image (Ψ_I) and text encoders (Ψ_T), akin to CLIP score [8]. *Diversity* (D) measures the average dissimilarity of generated views $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ from a reference view $\mathbf{y}$, reflecting how distinct the generated images are from the reference view, emphasizing creative variations. The diversity is computed by averaging the cosine similarity of each generated view with the reference view using CLIP image encoders.

$$D = \frac{1}{N}\sum_{n=1}^{N}\left[1 - cos(\Psi_I(\mathbf{y}), \Psi_I(\mathbf{x}^{(n)}))\right]. \tag{8}$$

Semantic variance (S_{Var}) quantifies the variance in semantic changes across views. This measures how similar the generated images are to a given text prompt, "`An image of {OBJECT}`". The semantic variance is calculated by averaging the cosine similarity between the CLIP text embedding of the prompt and the CLIP image embedding of each generated view, followed by measuring the variance of these values across views.

$$\begin{aligned}\bar{\mathrm{S}} &= \frac{1}{N}\sum_{n=1}^{N}\cos(\Psi_T(\texttt{text}), \Psi_I(\mathbf{x}^{(n)})), \\ \mathrm{S}_{Var} &= \frac{1}{N}\sum_{n=1}^{N}(\cos(\Psi_T(\texttt{text}), \Psi_I(\mathbf{x}^{(n)})) - \bar{\mathrm{S}})^2.\end{aligned} \tag{9}$$

The CD score is then computed as the ratio of diversity to semantic variances across views:

$$\text{CD Score} = D/\mathrm{S}_{Var}. \tag{10}$$

We note that the CD score is reference-free, *i.e.*, it does not require any ground truth images to measure the score.

4. Experiments

Due to space constraints, we provide detailed information regarding implementation details and baselines in Appendix. **Dataset.** Following [17–19], we used the Google Scanned Object (GSO) [6] dataset, adopting the same data split as in [19], for our evaluation. In addition, we utilized Internet-collected images, including those curated by [19], to assess the generation ability for complex objects or scenes.

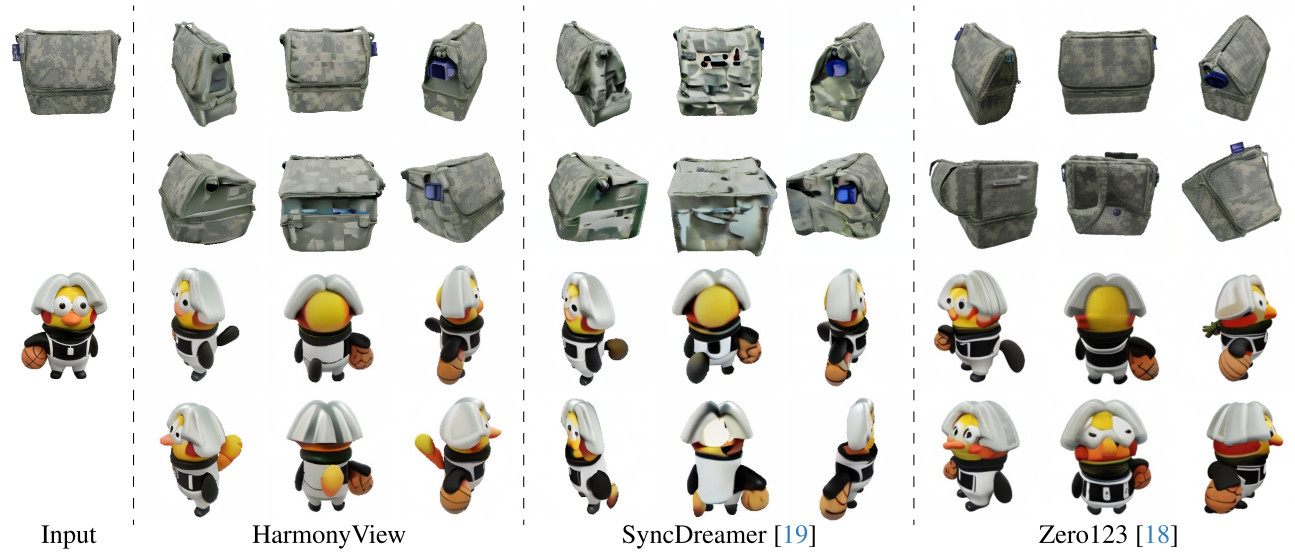

Figure 4. **Novel-view synthesis comparison.** HarmonyView generates plausible novel views while preserving coherence across views.

Method	PSNR↑	SSIM↑	LPIPS↓	E_{flow}↓	CD↑
Realfusion [22]	15.26	0.722	0.283	-	-
Zero123 [18]	18.98	0.795	0.166	3.820	0.628
SyncDreamer [19]	20.19	0.819	0.140	2.071	0.717
HarmonyView	**20.69**	**0.825**	**0.133**	**1.945**	**0.792**

Table 2. **Novel-view synthesis on GSO [6] dataset.** We report PSNR, SSIM, LPIPS, E_{flow}, and CD score.

Methods	CD↑	User Likert Score (1-5)↑		
		Quality	Consistency	Diversity
Zero123 [18]	0.752	3.208	3.167	2.854
SyncDreamer [19]	0.722	3.417	3.208	2.708
HarmonyView	**0.804**	**3.958**	**3.479**	**3.813**

Table 3. **Novel-view synthesis on in-the-wild images.** We report the CD score and 5-scale user Likert score, assessing quality, consistency, and diversity. Notably, the CD score shows strong alignment with human judgments. The test images are collected by [19].

Tasks and metrics. For the novel-view synthesis task, we used three standard metrics – PSNR, SSIM [48], LPIPS [58] – to measure sample quality compared to GT images. We measured diversity using the CD score. As a multi-view coherency metric, we propose E_{flow}, which measures the ℓ_1 distance between optical flow estimates from RAFT [43] for both GT and generated images. For the single-view 3D reconstruction task, we used Chamfer distance to evaluate point-by-point shape similarity and volumetric IoU to quantify the overlap between reconstructed and GT shapes.

4.1. Comparative Results

Novel-view synthesis. Table 2 shows the quantitative results for novel-view synthesis on the GSO [6] dataset. Here, HarmonyView outperforms state-of-the-art methods across all metrics. We confirm that HarmonyView generates images of superior quality, as indicated by PSNR, SSIM and LPIPS. It particularly excels in achieving multi-view coherency (indicated by E_{flow}) and generating diverse views that are faithful to the semantics of the input view (indicated by CD score). In Fig. 4, we present the qualitative results. Zero123 [18] produces multi-view incoherent images or implausible images, *e.g.*, eyes on the back. SyncDreamer [19] generates images that lack visual similarity to the input view or contain deficiencies, *e.g.*, flatness or hole on the back. In contrast, HarmonyView generates diverse yet plausible multi-view images while maintaining geometric coherence across views. In Table 3, we examine novel-view synthesis methods on in-the-wild images curated by [19]. For evaluation, we use CD score and user Likert ratings (1 to 5) along three criteria: quality, consistency, and diversity. While SyncDreamer [19] excels in quality and consistency scores when compared to Zero123 [18], Zero123 performs better in diversity and CD score. Notably, HarmonyView stands out with the highest CD score and superior user ratings. This suggests that HarmonyView effectively produces visually pleasing, realistic, and diverse images while being coherent across multiple views. The correlation between the CD score and the diversity score underscores the efficacy of the CD score in capturing the diversity of generated images.

3D reconstruction. In Table 4, we quantitatively compare our approach against various other 3D generation methods [13, 17–19, 22, 25, 29]. Both our method and SDS-free methods [18, 19] utilize NeuS [47], a neural reconstruction method for converting multi-view images into 3D shapes. To achieve faithful reconstruction of 3D mesh that aligns well with ground truth, the generated multi-view images should be geometrically coherent. Notably, HarmonyView achieves the best results by a significant margin in both Chamfer distance and volumetric IoU metrics, demonstrating the proficiency of

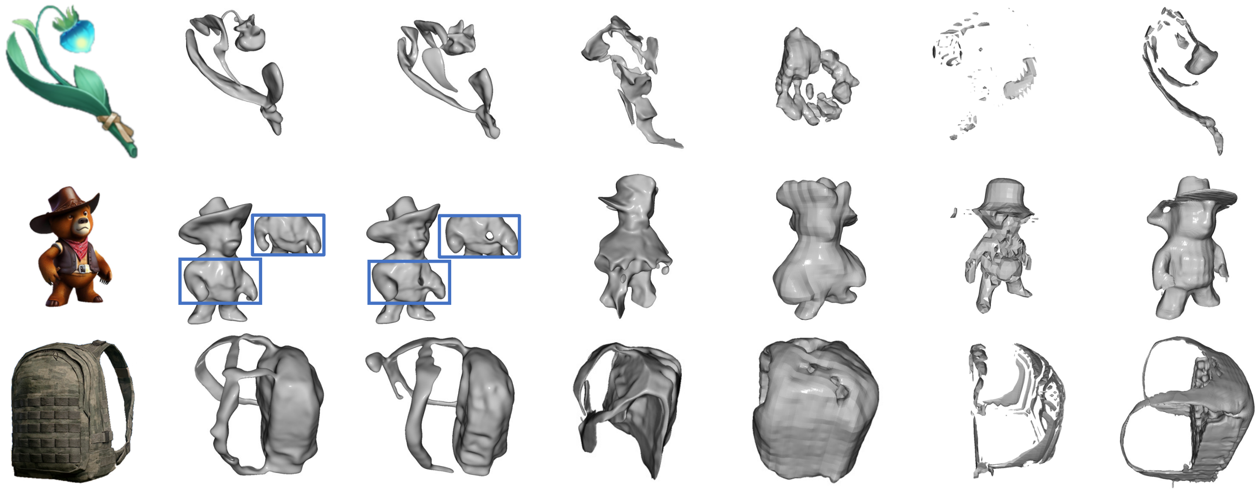

Figure 5. **3D reconstruction comparison.** HarmonyView stands out in creating high-quality 3D meshes where other often fails. HarmonyView, SyncDreamer [19], and Zero123 [18] use the vanilla NeuS [47] for 3D reconstruction.

Method	Chamfer Dist.↓	Volume IoU↑
Realfusion [22]	0.0819	0.2741
Magic123 [29]	0.0516	0.4528
One-2-3-45 [17]	0.0629	0.4086
Point-E [25]	0.0426	0.2875
Shap-E [13]	0.0436	0.3584
Zero123 [18]	0.0339	0.5035
SyncDreamer [19]	0.0261	0.5421
HarmonyView	**0.0187**	**0.6401**

Table 4. **3D reconstruction on GSO [6] dataset.** HarmonyView demonstrates substantial improvements over competitive baselines.

Method	s	s_1	s_2	PSNR↑	SSIM↑	LPIPS↓	E_{flow}↓	CD↑
Baseline (Eq. (3))	0.5	-	-	20.55	0.822	0.137	2.074	0.685
	1.0	-	-	20.19	0.819	0.140	2.071	0.717
	1.5	-	-	19.76	0.814	0.146	2.011	0.711
Ours (Eq. (7))	-	0.0	1.0	20.32	0.822	0.141	2.136	0.764
	-	1.0	1.0	20.55	0.824	0.135	2.009	0.772
	-	3.0	1.0	20.73	0.825	0.132	1.950	0.737
	-	2.0	0.0	**21.03**	**0.828**	**0.128**	2.146	0.668
	-	2.0	0.6	20.90	0.827	0.130	1.996	0.770
	-	2.0	0.8	20.80	0.826	0.131	2.009	0.774
	-	2.0	1.2	20.56	0.824	0.135	1.996	0.760
	-	**2.0**	**1.0**	20.69	0.825	0.133	**1.945**	**0.792**

Table 5. **Guidance scale study on novel-view synthesis.** We compare two instantiations of multi-view diffusion guidance: Eq. (3) and Eq. (7). Our approach consistently outperforms the baseline. Increasing s_1 tends to enhance PSNR, SSIM, and LPIPS, while higher s_2 tends to improve CD score. Notably, the combined effect of s_1 and s_2 synergistically improves E_{flow}.

HarmonyView in producing multi-view coherent images. We also present a qualitative comparison in Fig. 5. The results showcase the remarkable quality of HarmonyView. While competing methods often struggle with incomplete reconstructions (*e.g.*, Point-E, Shap-E), fall short in capturing small details (*e.g.*, Zero123), and show discontinuities (*e.g.*, SyncDreamer) or artifacts (*e.g.*, One-2-3-45), our method produces high-quality 3D meshes characterized by accurate geometry and a realistic appearance.

4.2. Analysis

Scale study. In Table 5, we investigate two instantiations of multi-view diffusion guidance with different scale configurations: baseline (Eq. (3)) and our approach (Eq. (7)). As s increases from 0.5 to 1.5 in the baseline method, E_{flow} (indicating *multi-view coherency*) and CD score (indicating *diversity*) show an increasing trend. Simultaneously, PSNR, SSIM, and LPIPS (indicating *visual consistency*) show a declining trend. This implies a trade-off between visual consistency and diversity. In contrast, our method involves parameters s_1 and s_2. We observe that increasing s_1 provides stronger guidance in aligning multi-view images with the input view, leading to direct improvements in PSNR, SSIM, and LPIPS. Keeping s_1 fixed at 2.0, elevating s_2 tends to yield improved CD score, indicating an enhanced diversity in the generated images. However, given the inherent conflict between consistency and diversity, an increase in s_2 introduces a trade-off. We note that our approach consistently outperforms the baseline across various configurations, striking a nuanced balance between consistency and diversity. Essentially, our decomposition provides more explicit control over those two dimensions, enabling a better balance. Additionally, the synergy between s_1 and s_2 notably enhances E_{flow}, leading to improved 3D alignment across multiple views.

Generalization to complex objects or scenes. Even in challenging scenarios, either with a highly detailed single object or multiple objects within a single scene, HarmonyView excels at capturing intricate details that SyncDreamer [19] might miss. The results are shown in Fig. 6. Our model well generates multi-view coherent images even in such scenar-

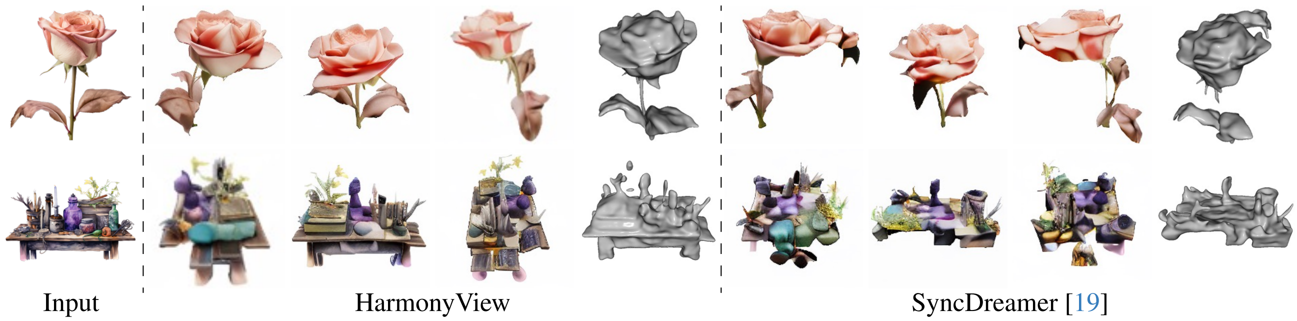

Figure 6. **3D reconstruction for complex object or scene.** HarmonyView successfully reconstructs the details, while SyncDreamer fails.

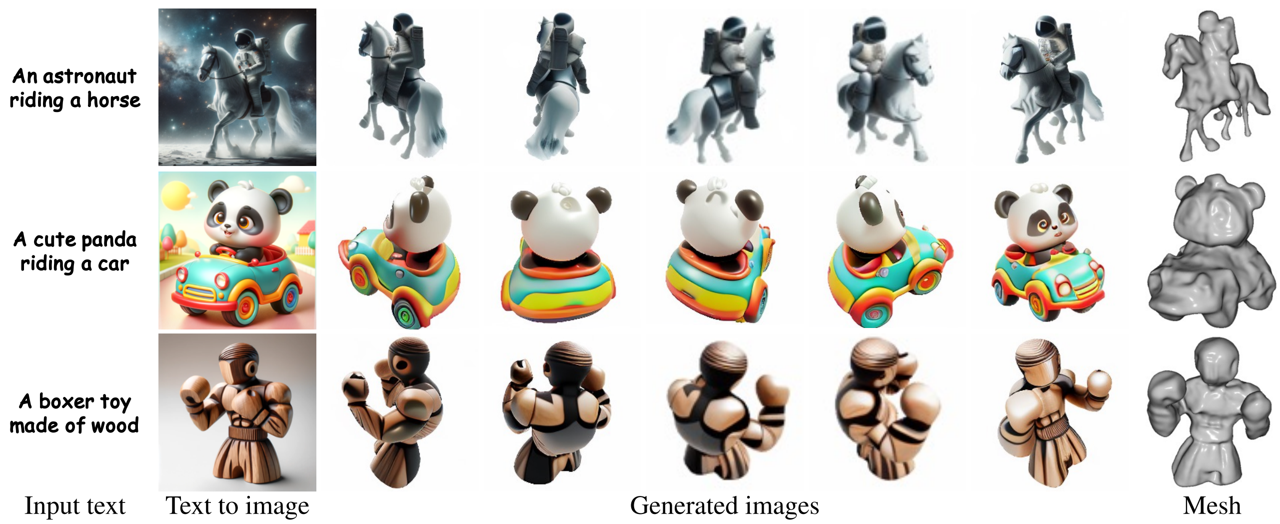

Figure 7. **Text-to-Image-to-3D.** HarmonyView, when combined with text-to-image frameworks [24, 31, 32], enables text-to-3D.

Method	DDIM Steps	PSNR↑	SSIM↑	LPIPS↓	E_{flow}↓	CD↑
SyncDreamer [19]	50	20.19	0.819	0.140	2.071	0.717
	200	20.20	0.823	0.140	2.009	0.727
HarmonyView	50	20.69	0.825	**0.133**	1.945	0.792
	200	**20.75**	**0.834**	**0.133**	**1.926**	**0.793**

Table 6. **Impact of sampling steps on novel-view synthesis.**

ios, enabling the smooth reconstruction of natural-looking meshes without any discontinuities.

Compatibility with text-to-image models. HarmonyView seamlessly integrates with off-the-shelf text-to-image models [31, 32]. These models convert textual descriptions into 2D images, which our model further transforms into high-quality multi-view images and 3D meshes. Visual examples are shown in Fig. 7. Notably, our model excels in capturing the essence or mood of the given 2D image, even managing to create plausible details for occluded parts. This demonstrates strong generalization capability, allowing it to perform well even with unstructured real-world images.

Runtime. HarmonyView generates 64 images (*i.e.*, 4 instances $\times$ 16 views) in only one minute, with 50 DDIM [39] sampling steps on an 80GB A100 GPU. Despite the additional forward pass through the diffusion model, HarmonyView takes less runtime than SyncDreamer [19], which requires about 2.7 minutes with 200 DDIM sampling steps. We also confirm that HarmonyView with 200 DDIM steps further improves the image quality, as shown in Table 6.

Additional results & analysis. Please see Appendix for more qualitative examples and analysis on the CD score, *etc*.

5. Conclusion

In this study, we have introduced HarmonyView, a simple yet effective technique that adeptly balances two fundamental aspects in a single-image 3D generation: *consistency* and *diversity*. By providing explicit control over the diffusion sampling process, HarmonyView achieves a harmonious equilibrium, facilitating the generation of diverse yet plausible novel views while enhancing consistency. Our proposed evaluation metric CD score effectively measures the diversity of generated multi-views, closely aligning with human evaluators' judgments. Experiments show the superiority of HarmonyView over state-of-the-art methods in both novel-view synthesis and 3D reconstruction tasks. The visual fidelity and faithful reconstructions achieved by HarmonyView highlight its efficacy and potential for various applications.

References

[1] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023. 2

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[3] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 3

[4] Rahul Dey and Vishnu Naresh Boddeti. Generating diverse 3d reconstructions from a single occluded face image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1547–1557, 2022. 3

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4

[6] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 5, 6, 7

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2, 5

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[11] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 3

[12] Yifan Jiang, Hao Tang, Jen-Hao Rick Chang, Liangchen Song, Zhangyang Wang, and Liangliang Cao. Efficient-3dim: Learning a generalizable single-image novel-view synthesizer in one day. *arXiv preprint arXiv:2310.03015*, 2023. 2

[13] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 6, 7

[14] Yash Kant, Aliaksandr Siarohin, Michael Vasilkovsky, Riza Alp Guler, Jian Ren, Sergey Tulyakov, and Igor Gilitschenski. invs: Repurposing diffusion inpainters for novel view synthesis. *arXiv preprint arXiv:2310.16167*, 2023. 2

[15] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3

[16] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261*, 2023. 2

[17] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 2, 3, 5, 6, 7

[18] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2, 3, 6, 7

[19] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1, 2, 3, 5, 6, 7, 8

[20] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 2

[21] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. Att3d: Amortized text-to-3d object synthesis. *arXiv preprint arXiv:2306.07349*, 2023. 3

[22] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 3, 6, 7

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3, 8

[25] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 6, 7

[26] Byeongjun Park, Hyojun Go, and Changick Kim. Bridging implicit and explicit geometric transformations for single-image view synthesis. *arXiv preprint arXiv:2209.07105*, 2022. 2

[27] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[28] Simon Prince, Adrian David Cheok, Farzam Farbiz, Todd Williamson, Nikolas Johnson, Mark Billinghurst, and Hirokazu Kato. 3d live: Real time captured content for mixed reality. In *Proceedings. International Symposium on Mixed and Augmented Reality*, pages 7–317. IEEE, 2002. 2

[29] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2, 3, 6, 7

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3, 8

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 8

[33] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 2

[34] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[35] Yukai Shi, Jianan Wang, He Cao, Boshi Tang, Xianbiao Qi, Tianyu Yang, Yukun Huang, Shilong Liu, Lei Zhang, and Heung-Yeung Shum. Toss: High-quality text-guided novel view synthesis from a single image. *arXiv preprint arXiv:2310.10644*, 2023.

[36] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3

[37] Ayan Sinha, Asim Unmesh, Qixing Huang, and Karthik Ramani. Surfnet: Generating 3d shape surfaces using deep residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6040–6049, 2017. 2

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 8

[40] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. *arXiv preprint arXiv:2306.07881*, 2023. 2, 3

[41] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 3

[42] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint arXiv:2307.01097*, 2023. 2, 3

[43] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 6

[44] Shubham Tulsiani, Abhishek Kar, Joao Carreira, and Jitendra Malik. Learning category-specific deformable 3d models for object reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):719–731, 2016. 2

[45] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 3

[46] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 2

[47] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3, 6, 7

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[49] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2, 3

[50] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2, 3

[51] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. 2

[52] Zhenzhen Weng, Zeyu Wang, and Serena Yeung. Zeroavatar: Zero-shot 3d avatar generation from a single image. *arXiv preprint arXiv:2305.16411*, 2023. 2, 3

[53] Sunu Wibirama, Paulus Insap Santosa, Putu Widyarani, Nanda Brilianto, and Wina Hafidh. Physical discomfort and

eye movements during arbitrary and optical flow-like motions in stereo 3d contents. *Virtual Reality*, 24(1):39–51, 2020. 2

[54] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021. 3

[55] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4479–4489, 2023. 3

[56] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. *arXiv preprint arXiv:2310.10343*, 2023. 2, 3

[57] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. *arXiv preprint arXiv:2310.03020*, 2023. 2, 3

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[59] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, and Jiajun Wu. Learning to reconstruct shapes from unseen classes. *Advances in neural information processing systems*, 31, 2018. 2

[60] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023. 2

[61] Zi-Xin Zou, Weihao Cheng, Yan-Pei Cao, Shi-Sheng Huang, Ying Shan, and Song-Hai Zhang. Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views. *arXiv preprint arXiv:2308.14078*, 2023. 2, 3